Analiza si vizualizarea datelor

Nicoleta ROGOVSCHI

nicoleta.rogovschi@parisdescartes.fr

Outline

- Definition and objectives
- Methods of data visualization
- Examples of applications
- The curse of dimensionality
- Classification of high-dimensional data
- Techniques of dimensions reducing

Why visualising data ?

- Better presentation of data =>

  Better Understanding / Analysis

- "Goal ....is to communicate information clearly and effectively through graphical means."
  - Friedman(2008)

Motivation

- Increasing computational power of computers
- Data avalability
- A lot of high dimensional data
- An additional component of the clustering techniques

Motivation

A good visualization technique should be applied even if:

- We have few a priori knowledge about the data or not at all
- Exploratory goals are vague
- The data are inhomogeneous and noisy

Motivation

- Nowadays high-dimensional data are everywhere
- Associated with different Machine Learning tasks such as : classification, "clustering" and regression

Results of the visualization

The results of the visualizations can be represented in the form of:

- Maps
- Graphics
- Dashboarding

Why dimension reduction?

- The number of features can be very large
  - Genomic data: expressions of genes
    - Thousands of variables
  - Image data : each pixel of an image
    - An image 64X64 = 4096 features
  - Text categorization: frequency of phrases (or words) in a document or web page
    - More than ten thousand features

Methods of data visualization

Methods of data visualization

There are many methods of visualization according to the type of treated data.

The data can be :

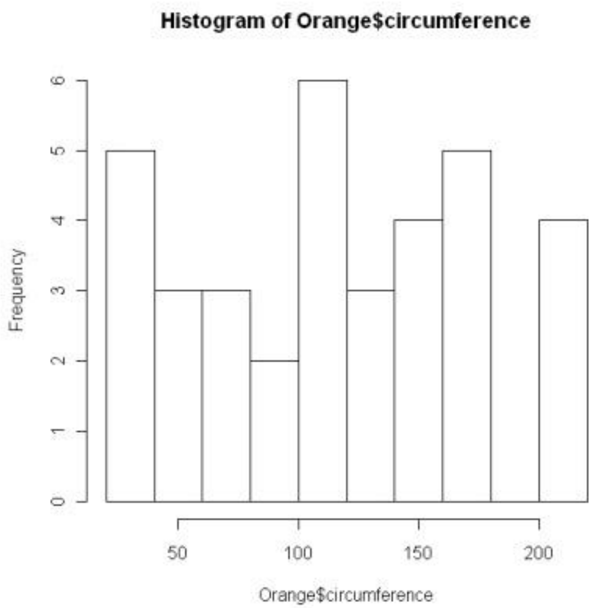
- Univariate
- Bivariate
- Multivariate

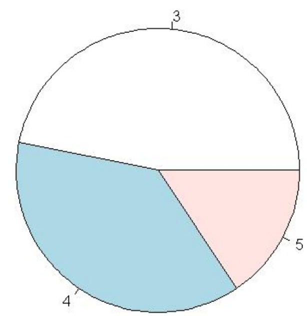
Univariate data

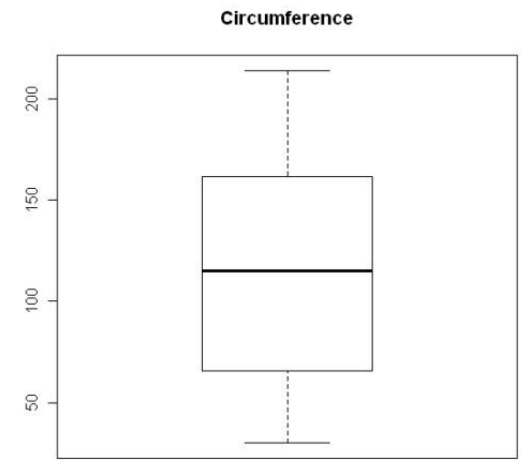
- Represents measurements of a variable
- Usually characterize a distribution
- Are represented by the following methods :
  - Histogram
  - Camambert (Pie chart)
  - Box plot

Univariate data (1D)

- Representation





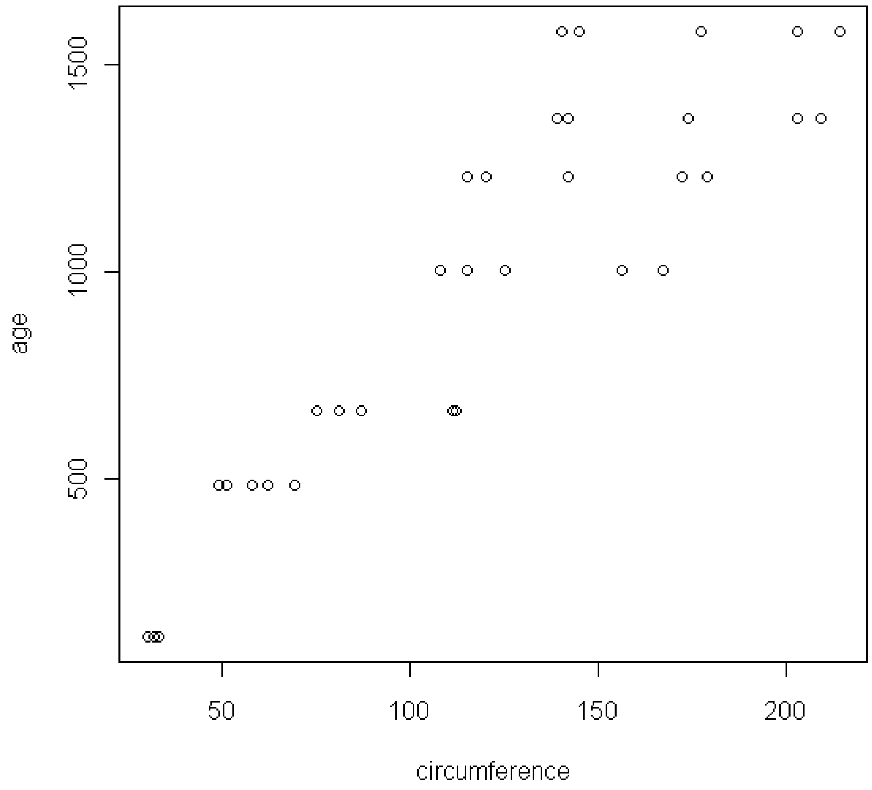


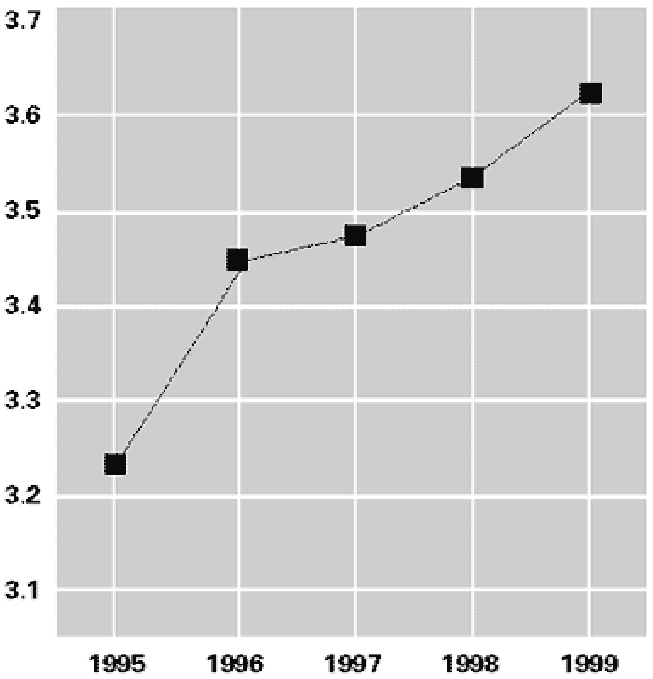
Bivariate data (2D)

- Are paired samples of two variables
- The variables are related
- Are represented by the following methods :
  - Scatter plot
  - Linear graphs

Bivariate data (2D)

- Representation





Multivariate data

- A multidimensional representation of multivariate data

- Are represented by the following methods :

  – Methods based icons

  – Pixel-based methods

  – Dynamic system in parallel coordinates

What is visualization?

What is visualization?

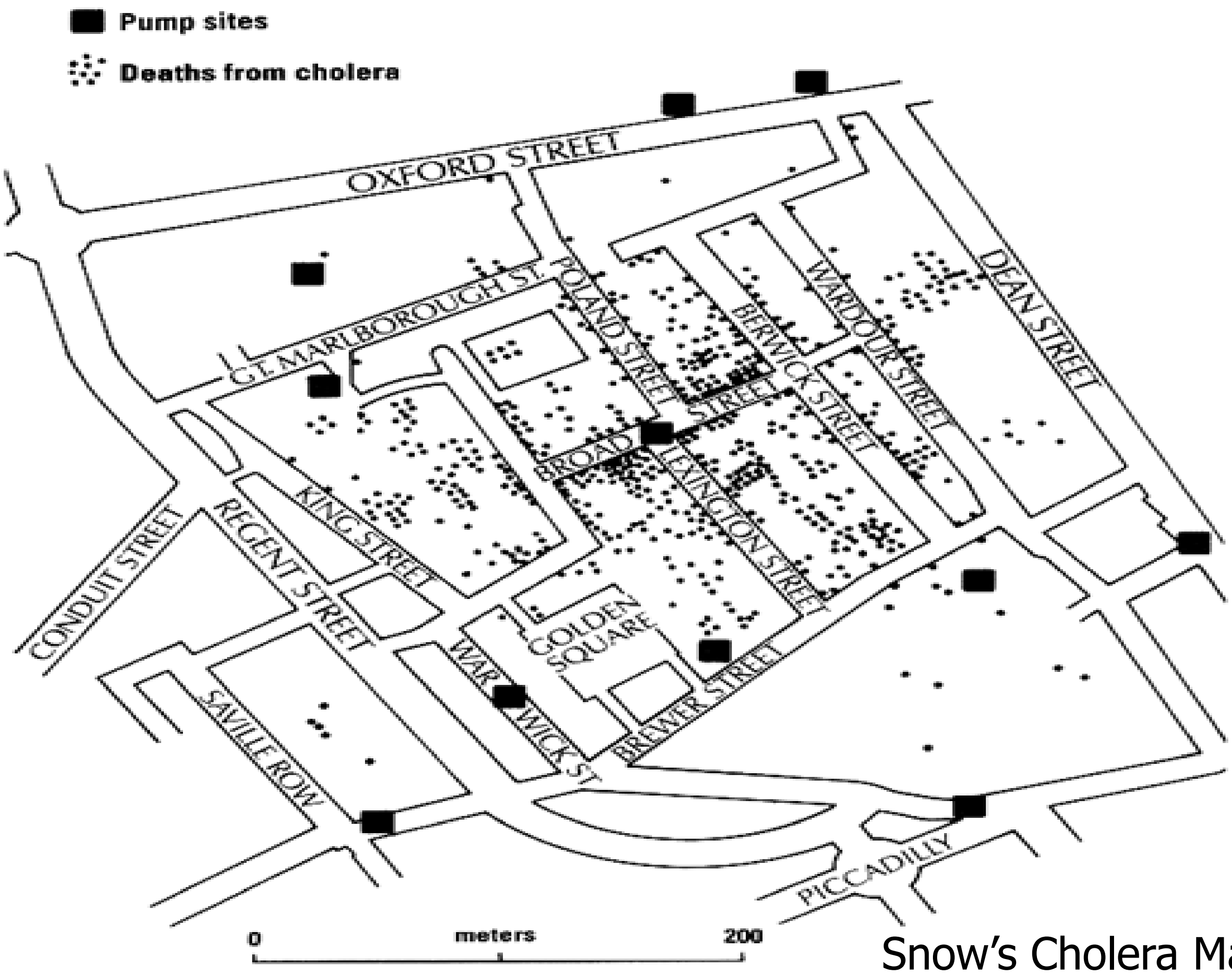
- Visualization is the process of **visual interpretation** or **graphical representation** of a dataset.

What is visualization?

- Medical investigations on patients

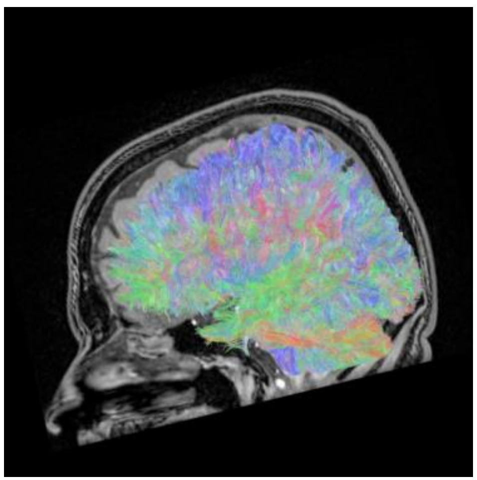
- 119 OE - pulse rate 26/07/2010 62 119 Diastolic blood pressure 26/07/2010 80 119 Systolic blood pressure 26/07/2010 120 201 Free T4 level 26/07/2010 17.8 201 Serum TSH level 26/07/2010 4.55 201 Serum calcium 26/07/2010 2.37 201 Serum inorganic phosphate 26/07/2010 1.17 201 Serum total protein 26/07/2010 64 201 Serum albumin 26/07/2010 41 201 Serum globulin 26/07/2010 23 201 Serum bilirubin level 26/07/2010 12 201 Serum alkaline phosphatase 26/07/2010 237 201 ALTSGPT serum level 26/07/2010 3 201 AST - aspartate transamSGOT 26/07/2010 24 201 International normalised ratio 26/07/2010 2 201 International normalised ratio 26/07/2010 2 580 Prostate specific antigen 26/07/2010 0.7 631 OE - pulse rate 26/07/2010 83 631 Diastolic blood pressure 26/07/2010 80 631 Systolic blood pressure 26/07/2010 133 634 Urine creatinine 26/07/2010 11.74 634 Urine microalbumin 26/07/2010 8.8 634 Urine albumin:creatinine ratio 26/07/2010 0.75 634 OE - weight 26/07/2010 84.35 634 OE - height 26/07/2010 176 634 Body Mass Index 26/07/2010 27.23 634 Diastolic blood pressure 26/07/2010 80 634 Systolic blood pressure 26/07/2010 135 786 Free T4 level 26/07/2010 12.6 786 Serum TSH level 26/07/2010 2.58 786 Plasma C reactive protein 26/07/2010 32 786 Serum calcium 26/07/2010 2.48 786 Serum inorganic phosphate 26/07/2010 0.93 786 Serum total protein 26/07/2010 71 786 Serum albumin 26/07/2010 41 786 Serum globulin 26/07/2010 30 786 Serum bilirubin level 26/07/2010 6 786 Serum alkaline phosphatase 26/07/2010 233 786 ALTSGPT serum level 26/07/2010 9 786 AST - aspartate transamSGOT 26/07/2010 18 786 Serum sodium 26/07/2010 137 786 Serum potassium 26/07/2010 4.1 786 Serum chloride 26/07/2010 103 786 Serum urea level 26/07/2010 4.7 786 Serum creatinine 26/07/2010 83 786 Total white blood count 26/07/2010 8.8 786 Red blood cell RBC count 26/07/2010 4.32 786 Haemoglobin estimation 26/07/2010 13 786 Haematocrit - PCV 26/07/2010 38.2 786 Mean corpuscular volume MCV 26/07/2010 88.4 786 Mean corpusc haemoglobinMCH 26/07/2010 30 786 Mean corpusc Hb conc MCHC 26/07/2010 34 786 Platelet count 26/07/2010 364 786 Neutrophil count 26/07/2010 6.6 786 Lymphocyte count 26/07/2010 1.5 786 Monocyte count 26/07/2010 0.5 786 Eosinophil count 26/07/2010 0.1 786 Erythrocyte sedimentation rate 26/07/2010 21 816 Serum vitamin B12 26/07/2010 348 816 Serum folate 26/07/2010 3.8 816 Total white blood count 26/07/2010 3.9 816 Red blood cell RBC count 26/07/2010 3.83 816 Haemoglobin estimation 26/07/2010 13.3 816 Haematocrit - PCV 26/07/2010 39.4 816 Mean corpuscular volume MCV 26/07/2010 103 816 Mean corpusc haemoglobinMCH 26/07/2010 34.8 816 Mean corpusc Hb conc MCHC 26/07/2010 33.8 816 Platelet count 26/07/2010 137 816 Neutrophil count 26/07/2010 2.4 816 Lymphocyte count 26/07/2010 1 816 Monocyte count 26/07/2010 0.4 816 Eosinophil count 26/07/2010 0.1 816 Body Mass Index 26/07/2010 28.99 816 OE - weight 26/07/2010 92.05 816 OE - height 26/07/2010 178.2 816 OE - pulse rate 26/07/2010 52 816 Diastolic blood pressure 26/07/2010 54 816 Systolic blood pressure 26/07/2010 107 856 Free T4 level 26/07/2010 16.4 856 Serum TSH level 26/07/2010 2.95 856 Serum calcium 26/07/2010 2.6 856 Serum inorganic phosphate 26/07/2010 0.82 856 Serum total protein 26/07/2010 72 856 Serum albumin 26/07/2010 47 856 Serum globulin 26/07/2010 25 856 Serum bilirubin level 26/07/2010 15 856 Serum alkaline phosphatase 26/07/2010 176 856 ALTSGPT serum level 26/07/2010 33 856 AST - aspartate transamSGOT 26/07/2010 23 856 Serum sodium 26/07/2010 141 856 Serum potassium 26/07/2010 4.8 856 Serum chloride 26/07/2010 102 856 Serum urea level 26/07/2010 6.3 856 Serum creatinine 26/07/2010 98 856 Total white blood count 26/07/2010 5.8 856 Red blood cell RBC count 26/07/2010 5.04 856 Haemoglobin estimation 26/07/2010 16 856 Haematocrit - PCV 26/07/2010 47.1 856 Mean corpuscular volume MCV 26/07/2010 93.4 856 Mean corpusc haemoglobinMCH 26/07/2010 31.9 856 Mean corpusc Hb conc MCHC 26/07/2010 34.1 856 Platelet count 26/07/2010 162 856 Neutrophil count 26/07/2010 3.2 856 Lymphocyte count 26/07/2010 2.1 856 Monocyte count 26/07/2010 0.3 856 Eosinophil count 26/07/2010 0.2 856 Erythrocyte sedimentation rate 26/07/2010 5 856 Diastolic blood pressure 26/07/2010 90 856 Systolic blood pressure 26/07/2010 150 1005 International normalised ratio 26/07/2010 3 1163 Serum sodium 26/07/2010 141 1163 Serum potassium 26/07/2010 5 1163 Serum chloride 26/07/2010 98 1163 Serum urea level 26/07/2010 16.8 1163 Serum creatinine 26/07/2010 159 1163 Glomerular filtration rate 26/07/2010 28 1818 Free T4 level 26/07/2010 13.6 1818 Serum TSH level 26/07/2010 1.43 1818 Total white blood count 26/07/2010 5.8 1818 Red blood cell RBC count 26/07/2010 4.65 1818 Haemoglobin estimation 26/07/2010 14.8 1818 Haematocrit - PCV 26/07/2010 43.1 1818 Mean corpuscular volume MCV 26/07/2010 92.6 1818 Mean corpusc haemoglobinMCH 26/07/2010 31.9 1818 Mean corpusc Hb conc MCHC 26/07/2010 34.5 1818 Platelet count 26/07/2010 205 1818 Neutrophil count 26/07/2010 4 1818 Lymphocyte count 26/07/2010 1.4 1818 Monocyte count 26/07/2010 0.3 1818 Eosinophil count 26/07/2010 0.1 1818 Erythrocyte sedimentation rate 26/07/2010 15 1818 Blood glucose level 26/07/2010 5.4 1818 Serum calcium 26/07/2010 2.48 1818 Serum inorganic phosphate 26/07/2010 0.89 1818 Serum total protein 26/07/2010 76 1818 Serum albumin 26/07/2010 43 1818 Serum globulin 26/07/2010 33 1818 Serum bilirubin level 26/07/2010 11 1818 Serum alkaline phosphatase 26/07/2010 287 1818 ALTSGPT serum level 26/07/2010 24 1818 AST - aspartate transamSGOT 26/07/2010 25 1818 Serum sodium 26/07/2010 138 1818 Serum potassium 26/07/2010 4.2 1818 Serum chloride 26/07/2010 105 1818 Serum urea level 26/07/2010 3.6 1818 Serum creatinine 26/07/2010 83 2714 Serum sodium 26/07/2010 134 2714 Serum potassium 26/07/2010 4.4 2714 Serum chloride 26/07/2010 104 2714 Serum urea level 26/07/2010 9.8 2714 Serum creatinine 26/07/2010 200 2714 Glomerular filtration rate 26/07/2010 29 3459 International normalised ratio 26/07/2010 2.5 3735 OE - pulse rate 26/07/2010 72 3735 Diastolic blood pressure 26/07/2010 96 3735 Systolic blood pressure 26/07/2010 169 4219 OE - pulse rate 26/07/2010 70 4219 Diastolic blood pressure 26/07/2010 71 4219 Systolic blood pressure 26/07/2010 108 4219 International normalised ratio 26/07/2010 3.3 4285 OE - pulse rate 26/07/2010 64 4285 Diastolic blood pressure 26/07/2010 90 4285 Systolic blood pressure 26/07/2010 150 4355 Diastolic blood pressure 26/07/2010 80 4355 Systolic blood pressure 26/07/2010 120 4511 International normalised ratio 26/07/2010 2.5 4763 International normalised ratio 26/07/2010 2.9 5111 International normalised ratio 26/07/2010 1.5

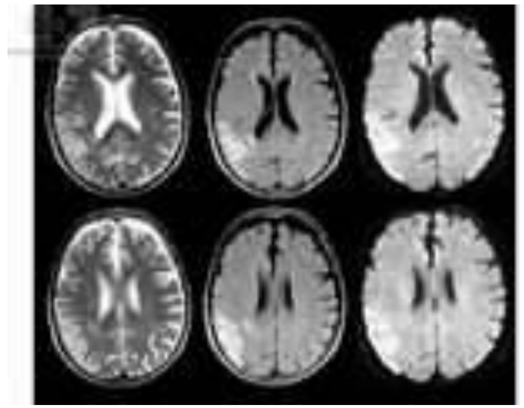
What is visualization?

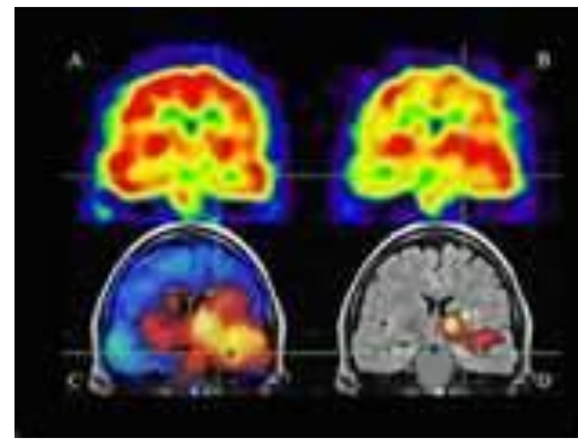


Snow's Cholera Map, 1855

What is visualization?







What is visualization?

What is visualization?

- Visualization is a branch of computer science involving the **processing, analysis and graphical representation of data** from diverse fields: social sciences, finance, medicine, entertainment, etc.

- There are two jurisdictions that are particularly requested in visualization: **computer graphics** and **statistics**.

What is visualization?

- It is important to know how to distinguish the areas of image processing, computer graphics and visualization.
- The **image processing is the study of 2D images to extract** information or to modify these characteristics.

What is visualization?

- The **computer graphics** allows to create images of any part using a computer, whether it be 2D images drawn by an artist or complex 3D scenes.

- The **visualization allows the exploration** of data represented in a visual form to help our understanding of the shown phenomenon.

What is visualization?

- One goal of visualization is to visually represent data that does not necessarily have a **natural geometric interpretation.**

Data Acquisition

The **acquisition of the raw data can be done in different** ways:

- by simulations (computer calculations)
- statistical surveys
- of historical databases
- of sensors of real measurements, etc.

Data Acquisition

Sources of errors

- Sampling is it accurate enough for us to be able to get the desired information? It should not be considered useless data that would only increase the calculations.
- The quantification is done with sufficient precision to be able to bring out the desired characteristics?

Filtering Data

Sources of errors

- Do we retain important and meaningful data? On the contrary, do we eliminate irrelevant data to the extraction of the desired characteristics?
- If we add data, the added data are representative of the rest?

High-dimensional data

- The increase in computing power and storage space of computers has led to an increase in the size of data sets
- Thus, several fields of science now are based on our ability to analyze and visualize high-dimensional data.

Curse of dimensionality

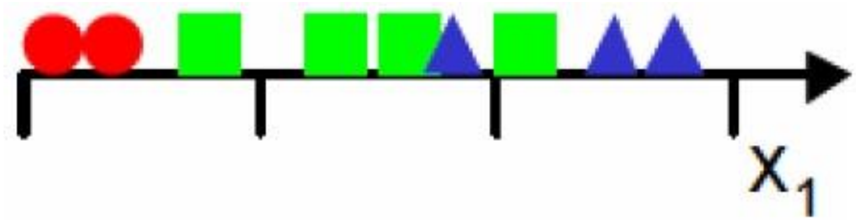
Curse of dimensionality

The curse of dimensionality

- A term introduced by Bellman in 1961
- Refers to the problem of the explosive increase in data volume associated with adding extra dimensions in a mathematical space.
- We will illustrate this problem with a simple example

Toy problem

- We have a 3-class pattern recognition problem
- We have available 9 observations 1D (along an axis)



- A simple approach would be to:
  - Divide the feature space into uniform bins
  - Compute the ratio of examples for each class at each bin and
  - For a new example, find its bin and choose the predominant class of the bin

Toy problem

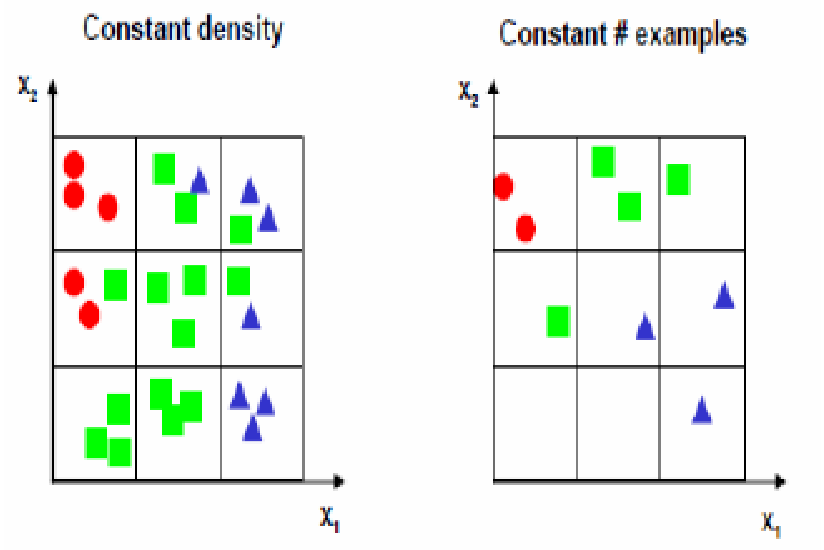
- For our example we start with one single feature and divide the axis into 3 segments. We observe that we have an average of 3 examples by region.
- Thereafter, we observe that there are too much overlap among the classes, so we decided to add a second feature to try to improve the class separability.

Toy problem (2D)

- If we add a 2nd dimension we pass from 3 cases (in 1D) to $3^2=9$ (in 2D).
- We have an another problem: do we maintain the density of examples per bin or do we keep the same number of example that was used in 1D?

Toy problem (2D)

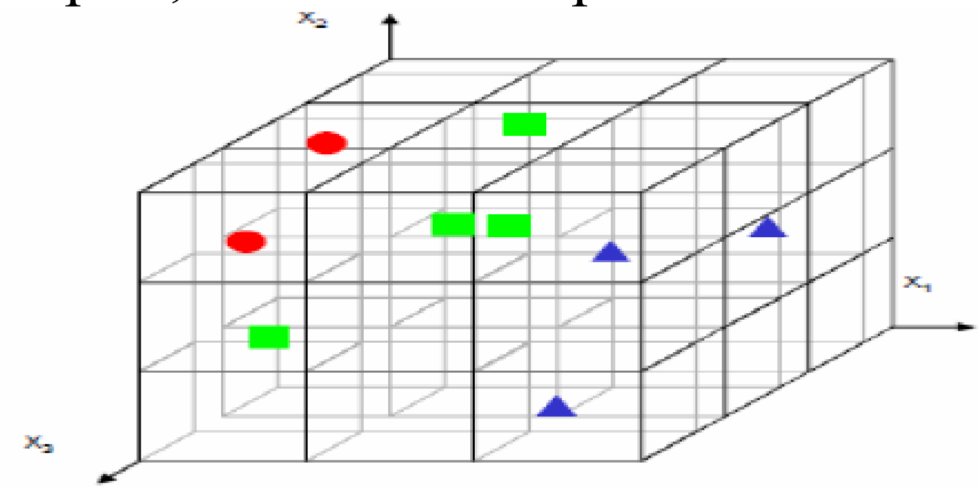
- Choosing to maintain the density increases the number of examples from 9 (in 1D) to 27 (in 2D)
- Choosing to maintain the number of examples results in a 2D scatter plot that is very sparse



Toy problem (3D)

If we add a 2nd dimension it gets worse :

- The number of bins grows to 3^3=27
- To keep the same density of examples the number of needed examples becomes 81
- For the same number of examples, the 3D scatter plot is almost empty



Curse of dimensionality

The approach performed on the toy example is ineffective :

- There are other approaches less affected by the curse of dimensionality, but the problem still exists

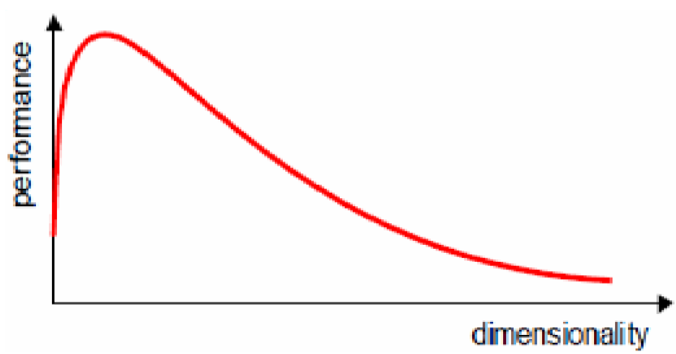
Curse of dimensionality

How can we beat the "curse of dimensionality "?

- By incorporating prior knowledge
- By providing increasing smoothness of the target function
- By reducing the dimensionality

Curse of dimensionality

- In practice, the curse of dimensionality means that, for a given sample size, there is a maximum number of variables beyond which the performance of our classifier will degrade rather than improve.



Curse of dimensionality

The curse of dimensionality generates several phenomena as:

- The concentration of the measurement
- Desertification of the space
- Depopulation of the center of hyper-volumes

Consequences

There are many consequences of the curse of the dimensionality:

1. Exponential growth in the number of examples required to mantain a given sampling density (For a density of **N** examples/bin and **D** dimensions, the total number of examples is $\mathbf{N^D}$)

Consequences

2. An exponential growth in the complexity of the target function (which estimates the density). To make a good learning, the target function requires denser sample points.

Consequences

3. For one dimension in the literature can be found a variety of density functions, but for high dimensions we have only the multivariate Gaussian density.

   In addition, for large values of D, we can treat the density only in a Gaussian simplified form.

Curse of dimensionality

- These findings suggest that we need special treatment to manipulate large data, which differs from that for low-dimensional data
- The same problems happens in other data distribution

Clustering High-Dimensional Data

Clustering high-dimensional data

Methods :

- **Subspace-clustering**: we are looking for clusteurs that exist in subspaces of the given high dimensional data
  - CLIQUE, ProClus and co-clustering approaches

- **Techniques of dimension reduction** : Construct a much lower dimensional space and search for clusters there (we may construct new dimensions by combining some dimensions in the original data).

Subspace-clustering

Methods of Subspace Clustering

- Subspace search methods: Search various subspaces to find clusters
  - "Bottom-up" approaches
  - "Top-down" approaches
- Clustering methods based on correlation
  - For instance : PCA based approaches
- Co-clustering methods
  - Optimization based methods (Cheng and Church, ISMB'2000)
  - Enumeration methods (Pei et al., ICDM'2003)

Subspace search methods

- **Subspace search methods**
- *“Bottom-up” approaches*
  - We start from low subspaces and search higher subspaces only they may be clusters in such subspaces
  - Various pruning techniques to reduce the number of higher subspaces to be searched
  - Ex. CLIQUE (Agrawal et al. 1998)
- *“Top-down” approaches*
  - We start from full space and we search smaller subspaces recursively
  - The subspace of the cluster can be determined by the local neighborhood
  - Ex. PROCLUS (Aggarwal et al. 1999): a *k*-medoid similar method

Methods based on correlation

- Clustering methods based on correlation : based on advanced correlation models

- Ex : PCA based approaches :
  - We apply PCA to generate a set of new uncorrelated methods
  - Then mine clusters in the new space or its subspaces

Methods of Co-clustering

- **Co-clustering** : Cluster both objects and attributes simultaneously (we treat objects and variables in symmetric way)
- Optimization based methods
  - We try to find a submatrix when she achieves the best significance as a co-cluster
  - Due to the cost in computation, greedy search is employed to find local optima co-clusters
- Enumeration methods
  - We use a tolerance threshold to specify the degree of noise allowed in the co-cluster to treat
  - The we try to enumerate all submatrices as co-clusters that satisfy the requirements

Techniques of dimension reduction

Dimension Reduction

- Data in a high-dimensional space are not uniformly distributed
- The reduction of dimension is a technique widely used to treat the "curse of dimensionality"

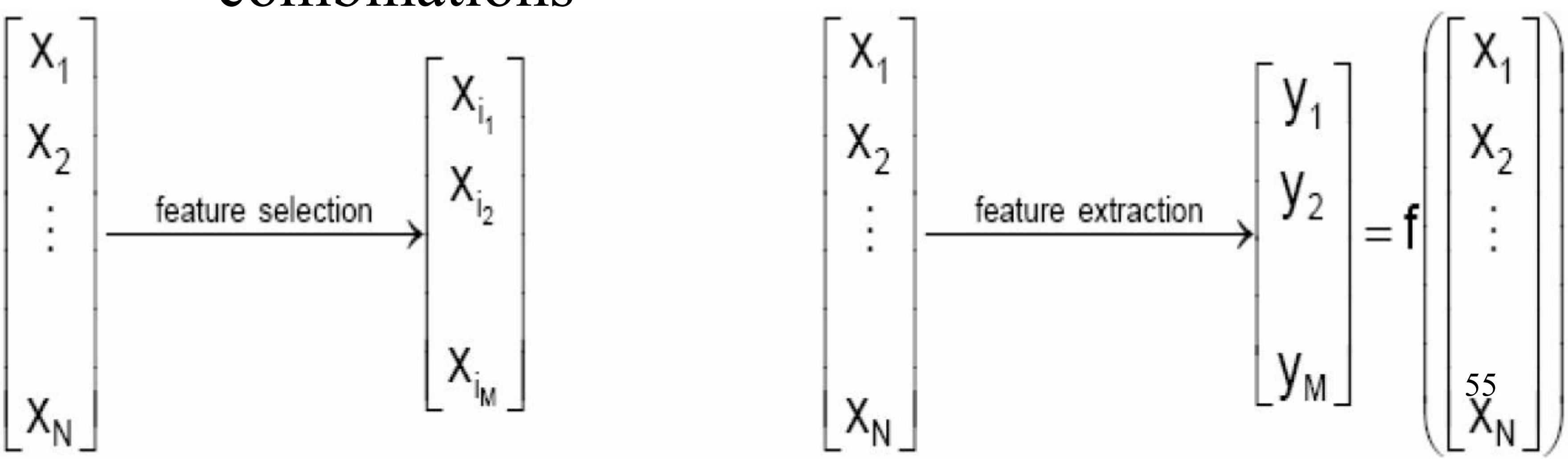
Dimension Reduction

There are a variety of techniques of dimension reduction:

- Linear vs. non-linear
- Deterministic vs. probabilistic
- Supervised vs. unsupervised

Dimension Reduction

- Dimension reduction : Methodologies
  - Feature selection: choosing a subset of all the feature
  - Feature extraction (« feature extraction ») : creating a subset of new features by combinations



Dimension reduction via Feature selection

Feature selection

- Definition :

Variable selection is a process to choose an optimal subset of relevant variables from a set of variables, according to a performance criterion.

We can ask three basic questions:

Q1: How to measure the relevance of the variables?
Q2: How to obtain the optimal subset?
Q3: Which optimality criterion to use?

Feature selection

- The answer to Q1 is to find a measure of relevance or evaluation criterion *J(X)* for quantify the importance of one variable or a combination of variables.

- Q2 refers to the problem of the choice of the procedure of research or creation of optimal subset of relevant variables.

- Q3 requires the definition of a stopping criterion of the research

Feature selection

Evaluation criteria

- For a classification problem, we test,
  for example, the discriminant quality of the system
  in the presence or absence of a variable.

- For a regression problem, we test
  rather the quality of prediction with respect to
  other variables.

Feature selection

An alternative is to use a search method of Branch & Bound type.

This search allows you to restrict the research and gives the optimal subset of variables, under the hypothesis of monotocity of the selection criterion *J(X)*.

The criterium is called monotonous if:

$$X_1 \subset X_2 \subset K \subset X_m \Rightarrow J(X_1) \leq J(X_2) \leq K \leq J(X_m)$$

where X_k is the subset containing the k selected variables.

Feature selection

Problem : most of the evaluation criteria are not monotonous

Use of sub-optimal methods: :

- *Sequential Forward Selection (SFS)*
- *Sequential Backward Selection (SBS)*
- *Bidirectional Selection (BS)*

Feature selection

- ***Sequential Forward Selection (SFS)***

Let X be the set of variables.

Initially the set of selected variables is empty.

At each step k: :

- We select the variable X_i that maximizes the criterion of evaluation $J(X_k)$

$$J(X_k) = \max_{x_i \in (X - X_{k-1})} J\left(X_{k-1} \cup \{x_i\}\right)$$

✓ ordered list of variables according to their importance

Feature selection

- *Sequential Backward Selection (SBS)*

We start from the full set of variables X and we perform by elimination :

at each step :

- The variable X_i the least important according to the evaluation criterion $J(X_k)$ is removed.

$$J(X_k) = \max_{x_i \in X_{k+1}} J\left(X_{k+1} - \{x_i\}\right)$$

✓ list of ordered variables according to their importance : The most relevant variables are then found in the last positions of the list.

Feature selection

The BS procedure performs the search in both directions (Forward and Backward) in a competitive manner.

The procedure stops in two cases:

(1) when one of the two directions has found the best subset of variables before reaching the middle of the search space

(2) when the two branches arriving in the middle.

Sets of selected variables found respectively by SFS and SBS are not equal because of their different principles of selection.

This method reduces the search time as the search is performed in both directions and stops when there is a solution regardless of the direction.

.

Feature selection

- *Stopping criterion*

✓ The optimal number of variables is unknown a priori, the use of a rule to control the selection/elimination of variables allows to stop the search when no variable is no longer enough informative.

✓ The stopping criterion is often defined as a combination of the search and the evaluation criteria.

✓ A heuristic often used is to calculate for different subsets of selected variables an estimation of the error of generalization by cross-validation.

✓ The subset of variables selected is that one which minimizes this generalization error.

Dimension reduction
via
Feature extraction

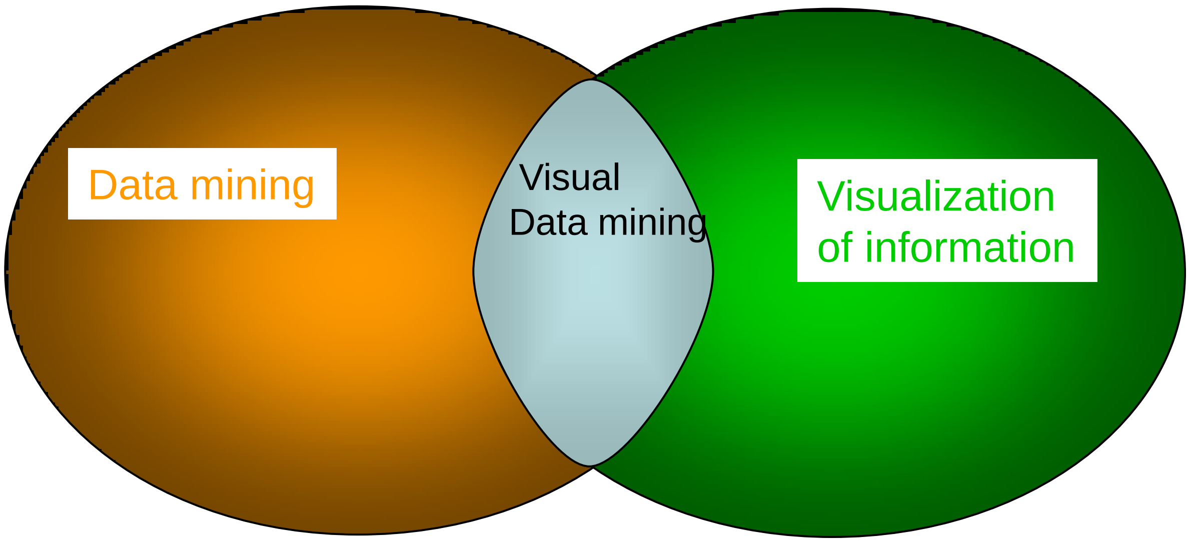
Dimension reduction via feature extraction

Two main types of methods :

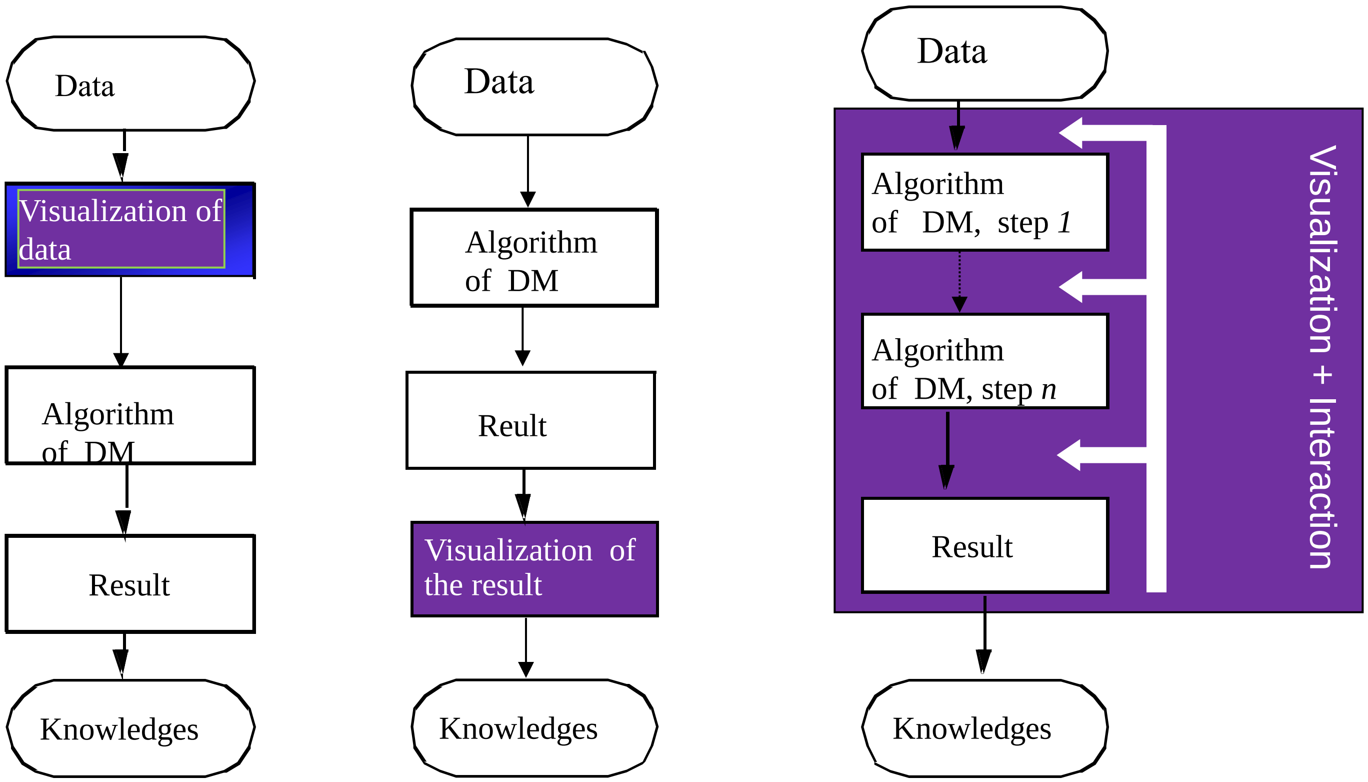
- **Linear Methods**
  - Principal Components Analysis (PCA)
  - Linear Discriminant Analysis (LDA)
  - Multi-Dimensional Scaling (MDS)
  - …
- **Non-Linear Methods**
  - Isometric feature mapping (Isomap)
  - Locally Linear Embedding (LLE)
  - Kernel PCA
  - Spectral clustering
  - Supervised methods (S-Isomap)
  - …

Visual data mining

Visual data mining



Visual data mining Diagram



Visualization software

- Graphviz
- Tulip
- Knime
- R
- ...

`www.KDnuggets.com/software/visualization.html`